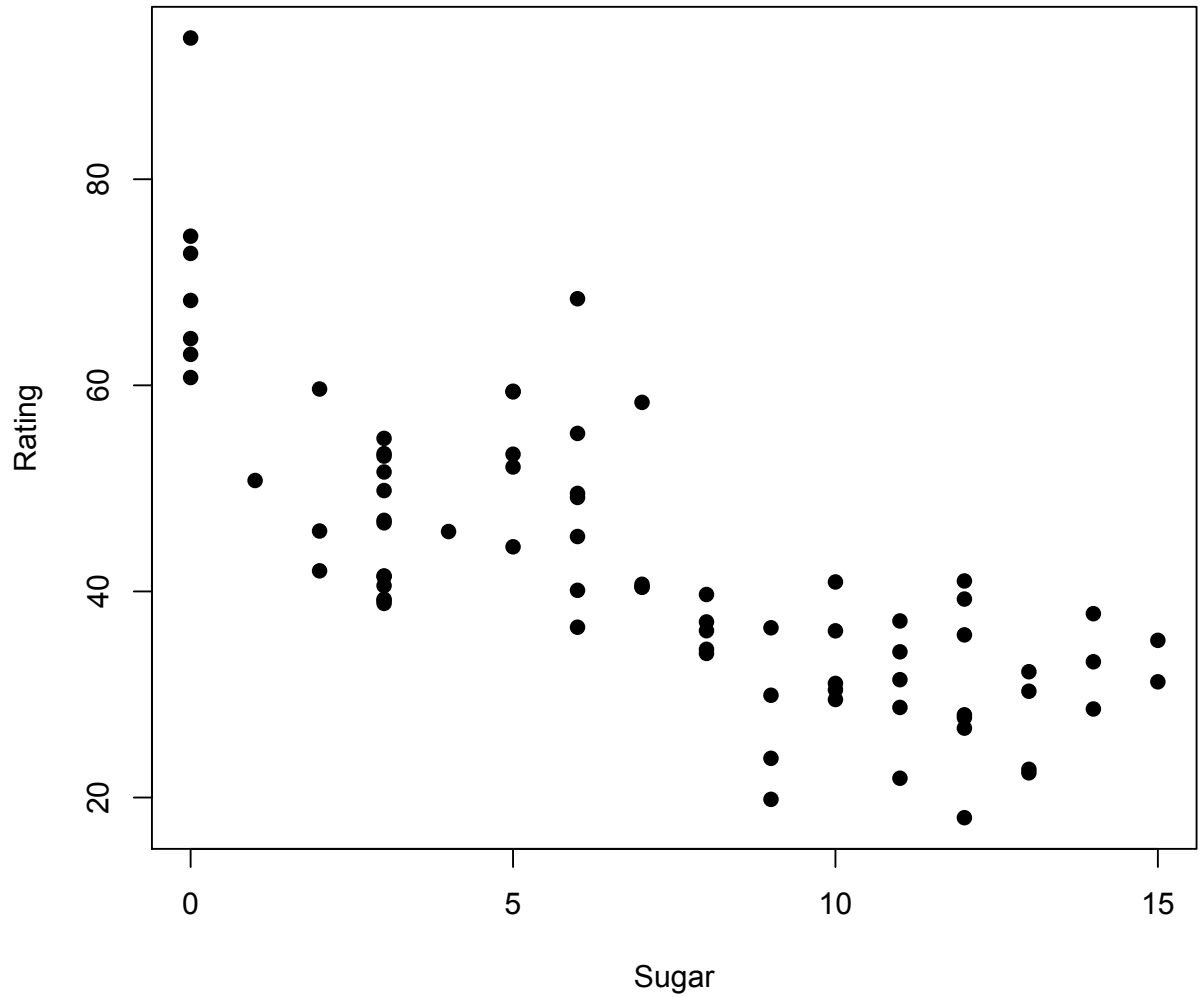
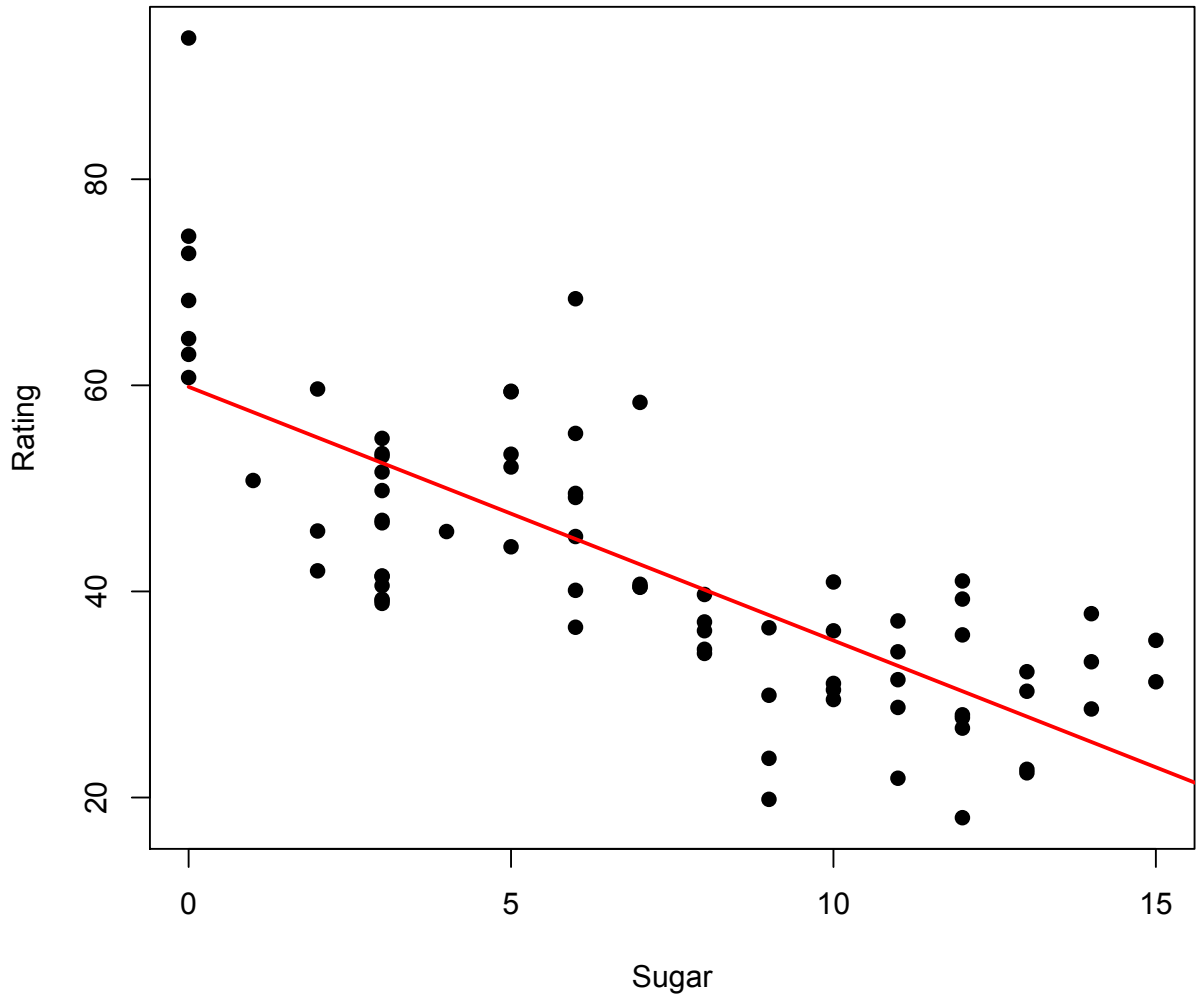
12 Simple Linear Regression

Material from Devore's book (Ed 8), and Cengagebrain.com

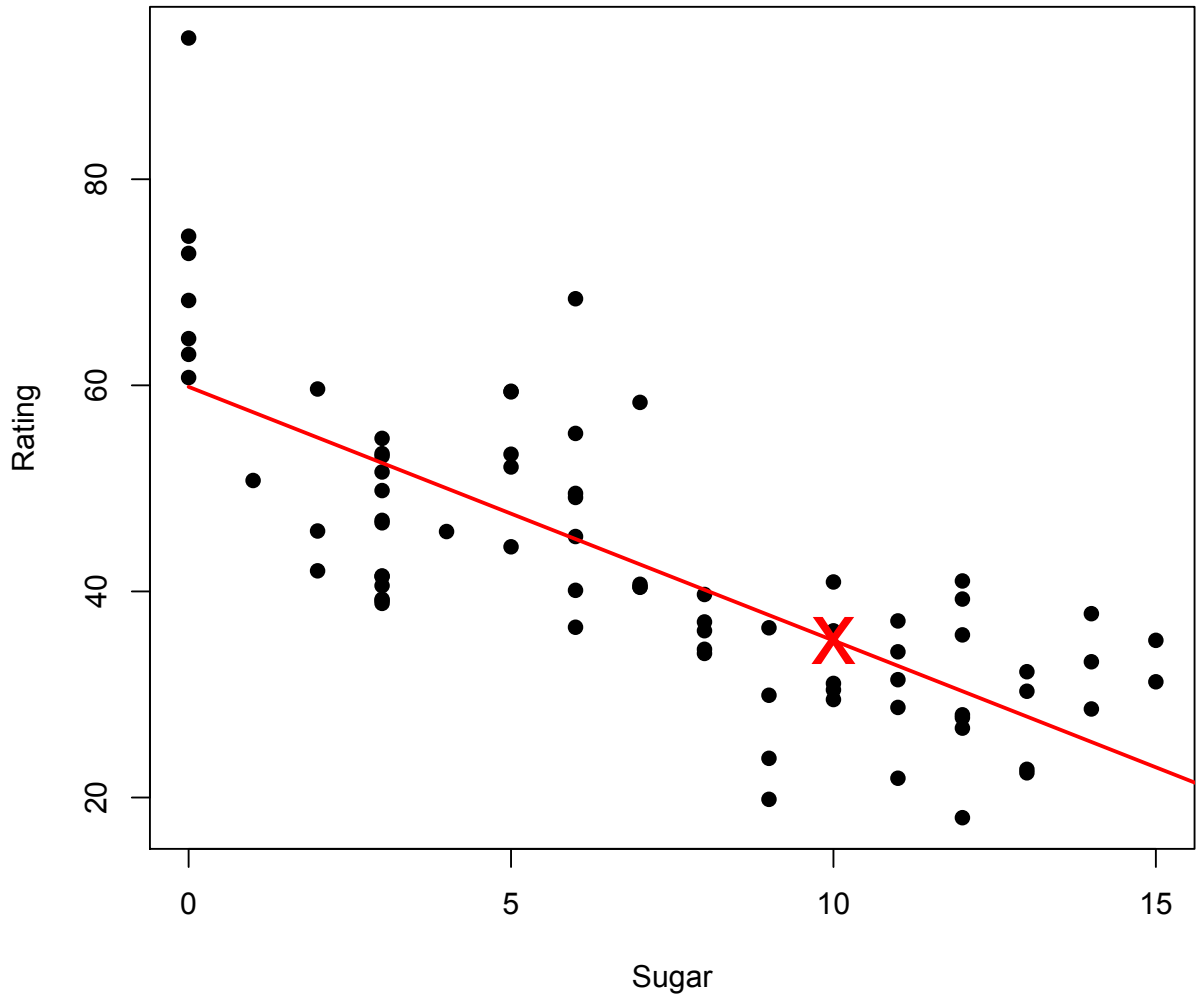
Simple Linear Regression



Simple Linear Regression



Simple Linear Regression



The Simple Linear Regression Model

The simplest deterministic *mathematical* relationship between two variables *x* and *y* is a linear relationship: $y = \beta_0 + \beta_1 x$.

The objective of this section is to develop an equivalent *linear probabilistic model*.

If the two (random) variables are probabilistically related, then for a fixed value of *x*, there is uncertainty in the value of the second variable.

So we assume $Y = \beta_0 + \beta_1 x + \varepsilon$, where ε is a random variable.

2 variables are related linearly "on average" if for fixed *x* the actual value of *Y* differs from its expected value by a random amount (*i.e.* there is random error).

A Linear Probabilistic Model

Definition The *Simple Linear Regression Model*

There are parameters β_0, β_1, and σ^2, such that *for any fixed value of the independent variable x, the dependent variable is a random variable* related to *x* through the **model equation**

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

The quantity ε in the model equation is the "error" -- a random variable, assumed to be symmetrically distributed with

$$E(\varepsilon) = 0 \text{ and } V(\varepsilon) = \sigma_\varepsilon^2 = \sigma^2$$

(no assumption made about the distribution of ε, yet)

A Linear Probabilistic Model

X: the **independent, predictor,** or **explanatory variable** (usually known). **NOT RANDOM.**

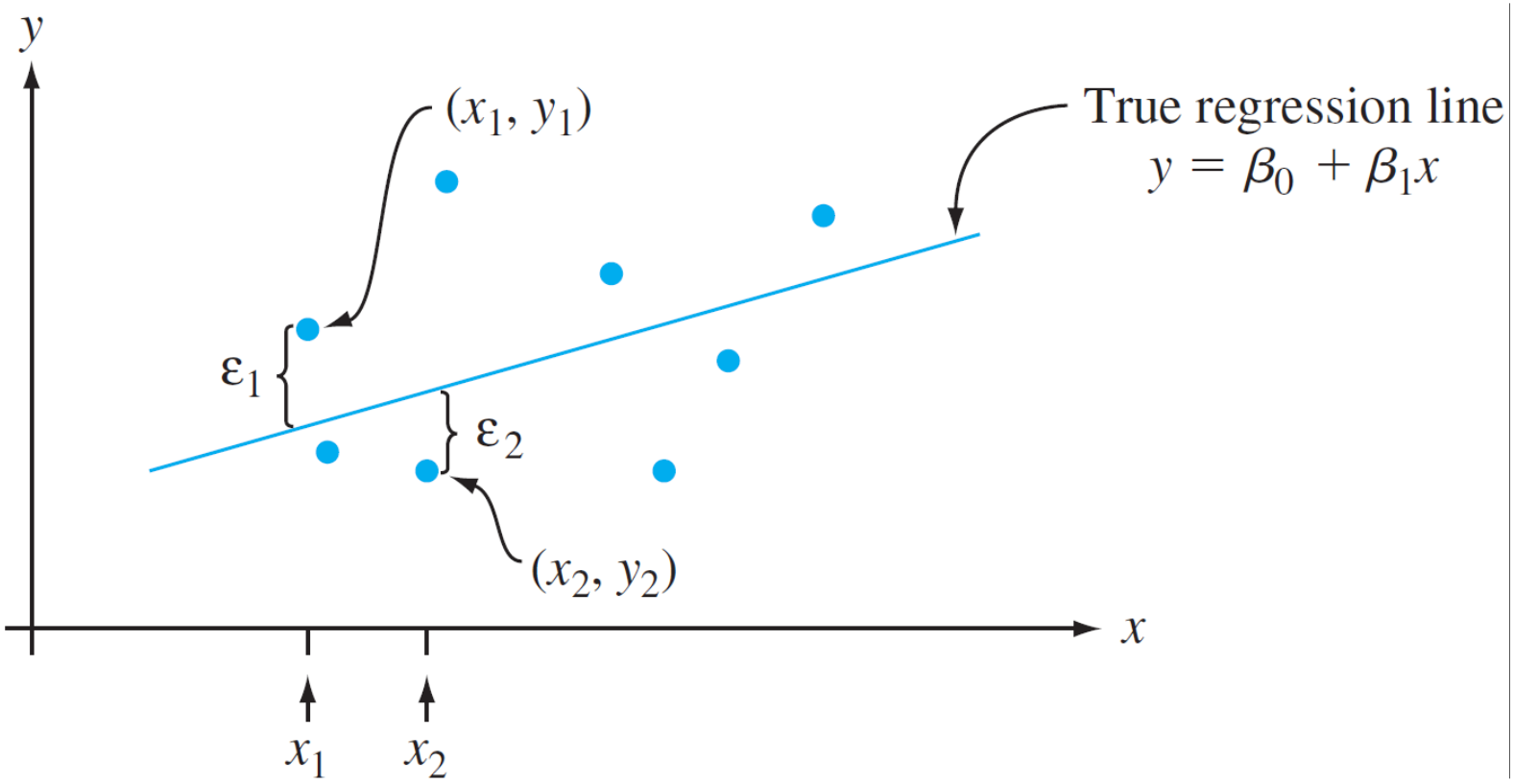
Y: The **dependent** or **response variable.** For **fixed *x*, *Y* will be random variable**.

ε: The **random deviation** or **random error term**. For **fixed *x*, *ε* will be random variable**.

What exactly does *ε do?*

A Linear Probabilistic Model

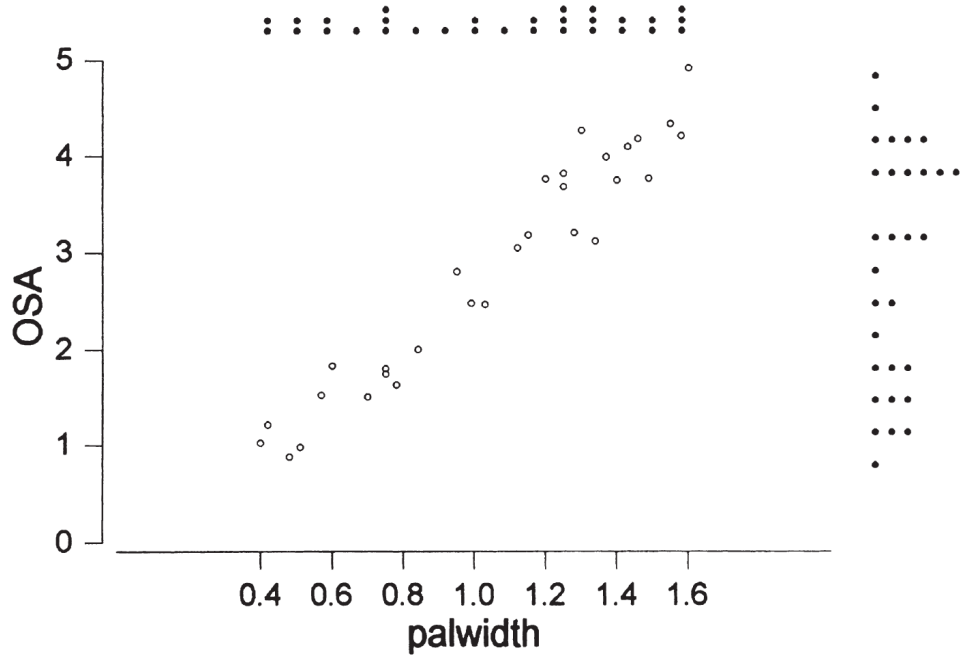
The points $(x_1, y_1), \ldots, (x_n, y_n)$ resulting from n independent observations will then be scattered about the true regression line:



A Linear Probabilistic Model

How do we know simple linear regression is appropriate?

- Theoretical considerations
- Scatterplots



A Linear Probabilistic Model

If we think of an entire population of (*x*, *y*) pairs, then $\mu_{Y|x*}$ is the *mean of all y values for which x = x**, and $\sigma^2_{Y|x*}$ is a measure of how much these values of *y* spread out about the mean value.

If, for example, *x* = age of a child and *y* = vocabulary size, then $\mu_{Y|5}$ is the average vocabulary size for all 5-year-old children in the population, and $\sigma^2_{Y|5}$ describes the amount of variability in vocabulary size for this part of the population.

A Linear Probabilistic Model

Interpreting parameters:

β_0 (the intercept of the true regression line):

The average value of Y when x is zero.

β_1 (the slope of the true regression line):

The ***expected (average)*** **change in *Y*** associated with a 1-unit increase in the value of *x*.

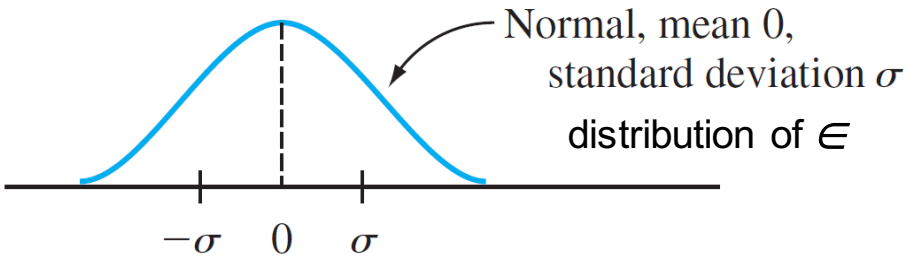
A Linear Probabilistic Model

What is $\sigma^2_{Y|x*}$? How do we interpret $\sigma^2_{Y|x}$?

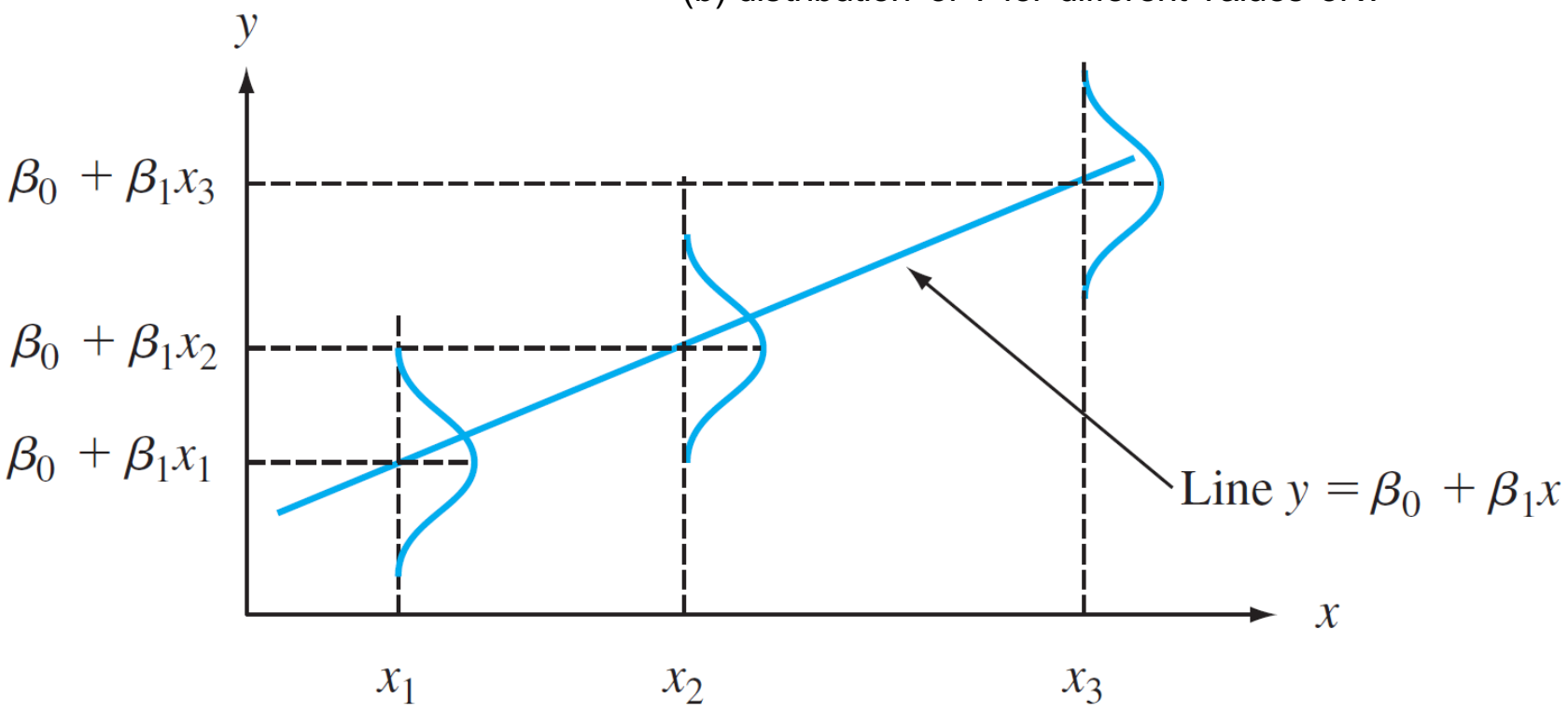
Homoscedasticity:

We assume the variance (amount of variability) of the distribution of *Y* values to be the same at each different value of fixed *x*. (*i.e.* homogeneity of variance assumption).

When errors are normally distributed…



(b) distribution of *Y* for different values of *x*



The variance parameter σ^2 determines the extent to which each normal curve spreads out about the regression line

A Linear Probabilistic Model

When σ^2 is small, an observed point (x, y) will almost always fall quite close to the true regression line, whereas observations may deviate considerably from their expected values (corresponding to points far from the line) when σ^2 is large.

Thus, this variance can be used to tell us how good the linear fit is

But how do we define “good”?

Estimating Model Parameters

The values of β_0, β_1, and σ^2 will almost never be known to an investigator.

Instead, sample data consists of n observed pairs

$$(x_1, y_1), \ldots, (x_n, y_n),$$

from which the model parameters and the true regression line itself can be estimated.

The data (pairs) are assumed to have been obtained independently of one another.

Estimating Model Parameters

Where

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ for } i = 1, 2, \ldots, n$$

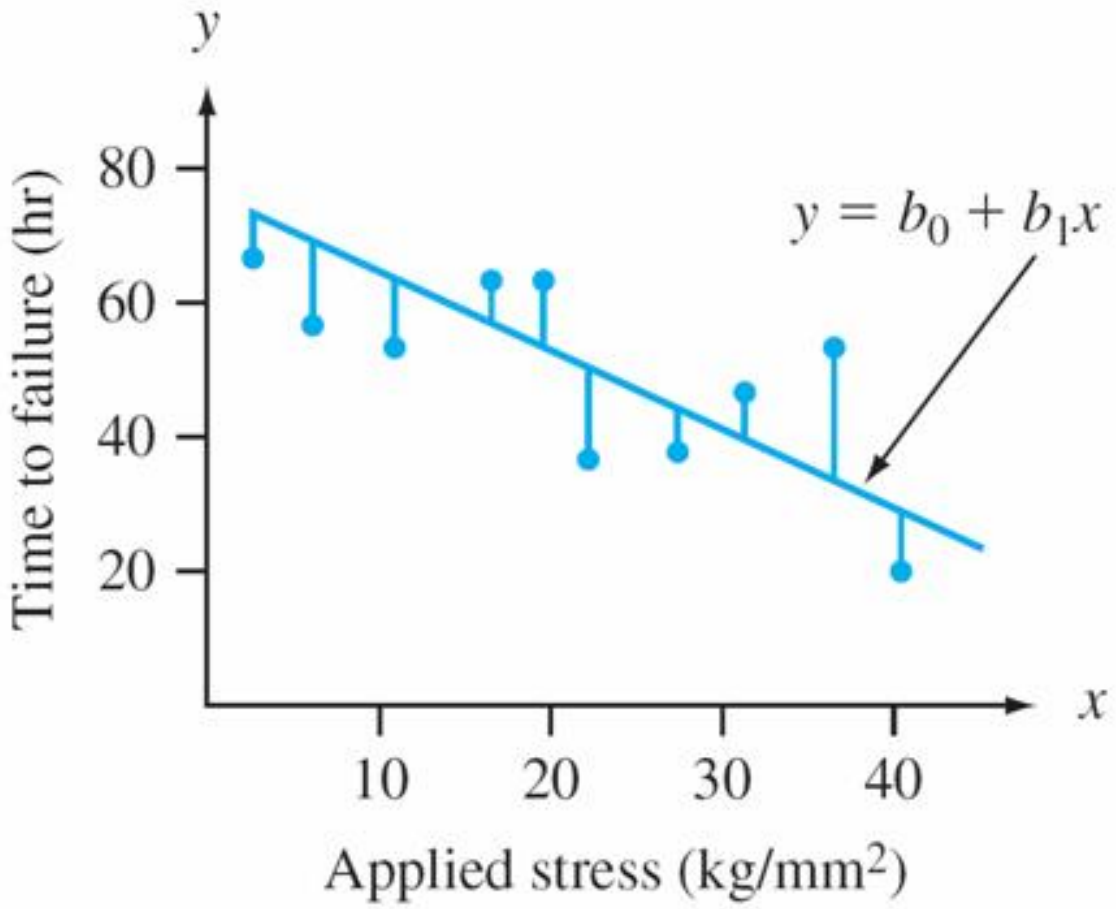
and the n deviations $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent r.v.'s.

($Y_1, Y_2, \ldots, Y_n$ are independent too, why?)

Estimating Model Parameters

The "best fit" line is motivated by the principle of **least squares**, which can be traced back to the German mathematician ***Gauss*** (1777–1855):

*A line provides the **best fit** to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.*



Estimating Model Parameters

The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, are called the **least squares estimates –** they are those values that minimize $f(b_0, b_1)$.

Estimating Model Parameters

The **fitted regression line** or **least squares line** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Estimating Model Parameters

The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Shortcut formulas for the numerator and denominator of $\hat{\beta}_1$ are

$$S_{xy} = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n \quad \textit{and} \quad S_{xx} = \Sigma x_i^2 - (\Sigma x_i)^2/n$$

(Typically columns for x_i, y_i, $x_i y_i$ and x_i^2 and constructed and then S_{xy} and S_{xx} are calculated.)

Estimating Model Parameters

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\Sigma y_i - \hat{\beta}_1 \Sigma x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

The computational formulas for S_{xy} and S_{xx} require only the summary statistics Σx_i, Σy_i, Σx_i^2 and $\Sigma x_i y_i$.

(Σy_i^2 will be needed shortly for the variance.)

Example (fitted regression line)

The *cetane number* is a critical property in specifying the ignition quality of a fuel used in a diesel engine.

Determination of this number for a biodiesel fuel is expensive and time-consuming.

The article "Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (*J. of Automobile Engr.*, 2009: 565–583) included the following data on x = iodine value (g) and y = cetane number for a sample of 14 biofuels (see next slide).

Example (fitted regression line)

cont'd

The iodine value (x) is the amount of iodine necessary to saturate a sample of 100 g of oil. The article's authors *fit the simple linear regression model to this data*, so let's do the same.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Calculating the relevant statistics gives

$\Sigma x_i = 1307.5, \qquad \Sigma y_i = 779.2,$

$\Sigma x_i^2 = 128{,}913.93, \qquad \Sigma x_i y_i = 71{,}347.30,$

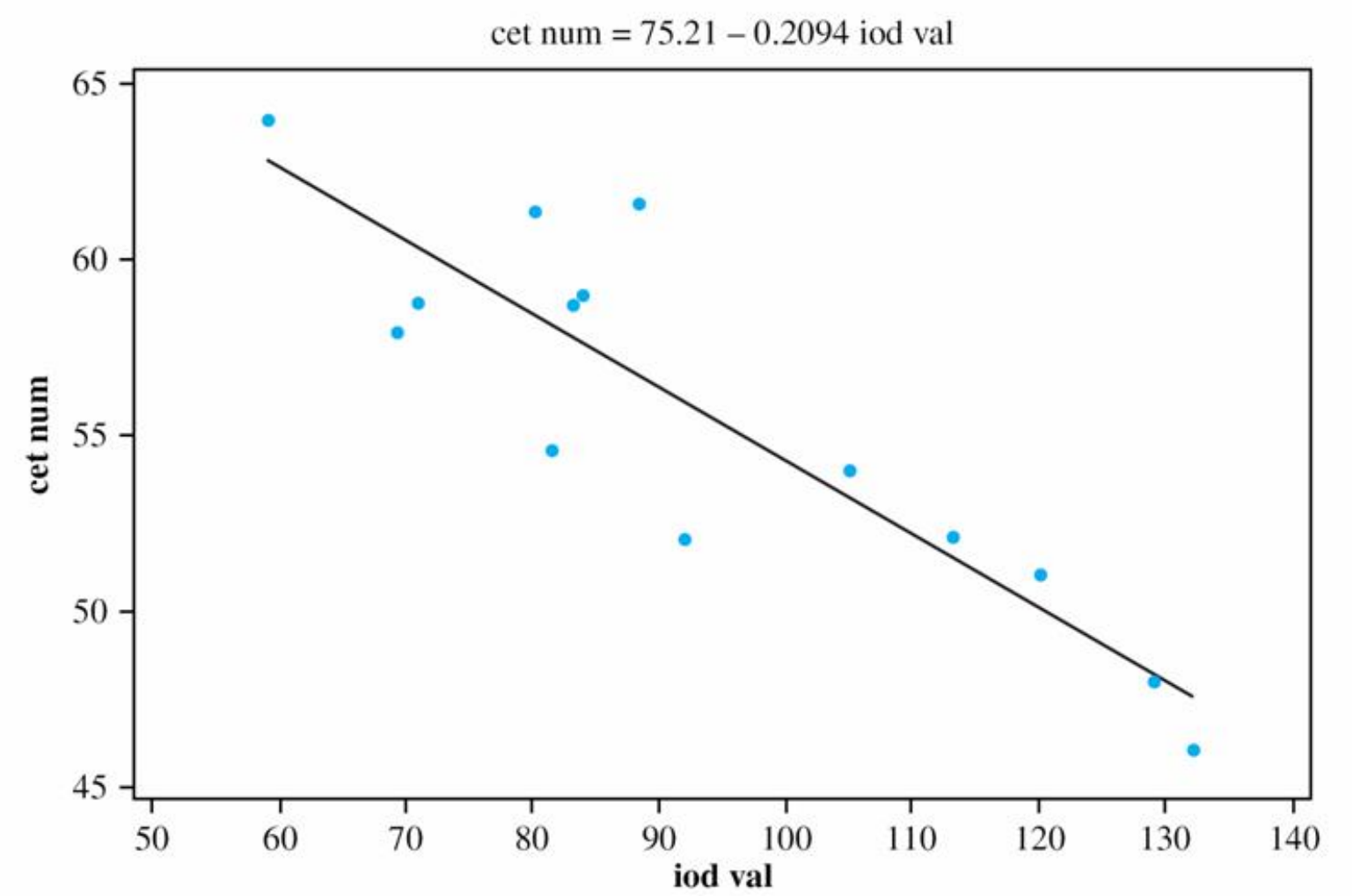
from which $S_{xx} = 128{,}913.93 - (1307.5)^2/14 = 6802.7693$

and $S_{xy} = 71{,}347.30 - (1307.5)(779.2)/14 = -1424.41429$

Example (fitted regression line)

cont' d

Scatter plot with the least squares line superimposed.



Fitted Values

Fitted Values:

The **fitted** (or **predicted**) **values** $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by substituting $x_1, \ldots, x_n$ into the equation of the estimated regression line:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$$

Residuals:

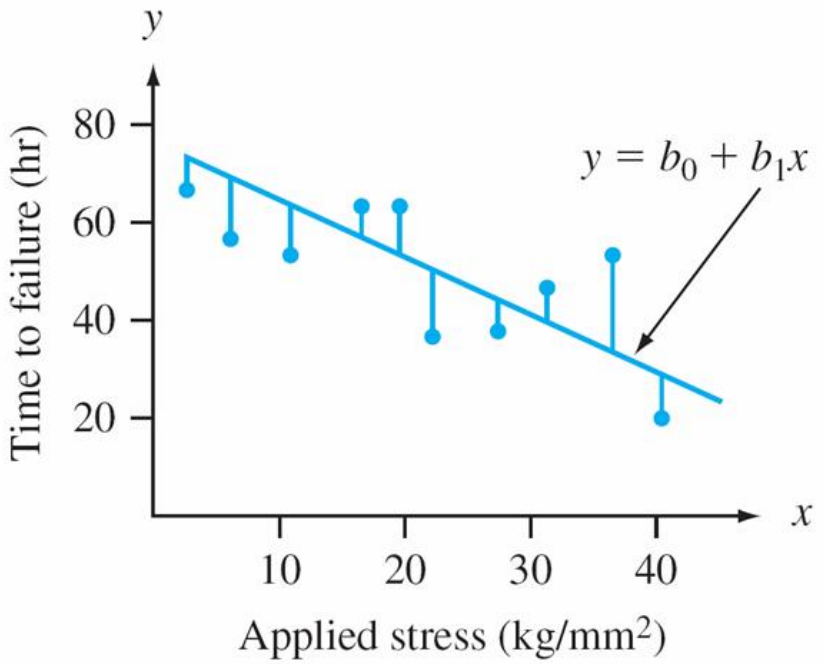
The differences $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$ between the observed and fitted *y* values.

Residuals are estimates of the true error – WHY?

Sum of the residuals

When the estimated regression line is obtained via the principle of least squares, *the sum of the residuals should in theory be zero*, if the error distribution is symmetric, since

$$\sum(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = n\overline{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\overline{x} = n\hat{\beta}_0 - n\hat{\beta}_0 = 0$$



Example (fitted values)

Suppose we have the following data on filtration rate (x) versus moisture content (y):

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Relevant summary quantities (*summary statistics*) are

$\Sigma x_i = 2817.9$, $\Sigma y_i = 1574.8$, $\Sigma x^2_i = 415{,}949.85$,

$\Sigma x_i\, y_i = 222{,}657.88$, and $\Sigma y^2_i = 124{,}039.58$,

From $S_{xx} = 18{,}921.8295$, $S_{xy} = 776.434$.

Calculation of residuals?

Example (fitted values)

cont' d

All predicted values (fits) and residuals appear in the accompanying table.

Obs	Filtrate	Moistcon	Fit	Residual
1	125.3	77.9	78.100	−0.200
2	98.2	76.8	76.988	−0.188
3	201.4	81.5	81.223	0.277
4	147.3	79.8	79.003	0.797
5	145.9	78.2	78.945	−0.745
6	124.7	78.3	78.075	0.225
7	112.2	77.5	77.563	−0.063
8	120.2	77.0	77.891	−0.891
9	161.2	80.1	79.573	0.527
10	178.9	80.2	80.299	−0.099
11	159.5	79.9	79.503	0.397
12	145.8	79.0	78.941	0.059
13	75.1	76.7	76.040	0.660
14	151.4	78.2	79.171	−0.971
15	144.2	79.5	78.876	0.624
16	125.0	78.1	78.088	0.012
17	198.8	81.5	81.116	0.384
18	132.5	77.0	78.396	−1.396
19	159.6	79.0	79.508	−0.508
20	110.7	78.6	77.501	1.099

Fitted Values

We interpret the fitted value as the value of *y* that we would predict or expect when using the estimated regression line with $x = x_i$; thus $\hat{y}_i$ is the ***estimated true mean*** for that population when $x = x_i$ (based on the data).

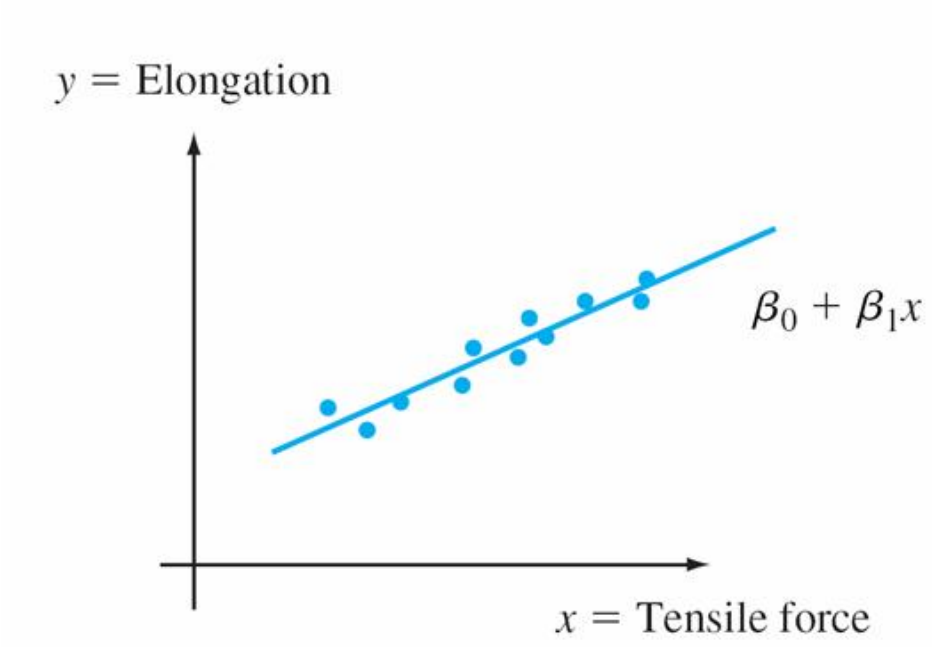
The residual $y_i - \hat{y}_i$ is a positive number if the point lies above the line and a negative number if it lies below the line. $(x_i, \hat{y}_i)$

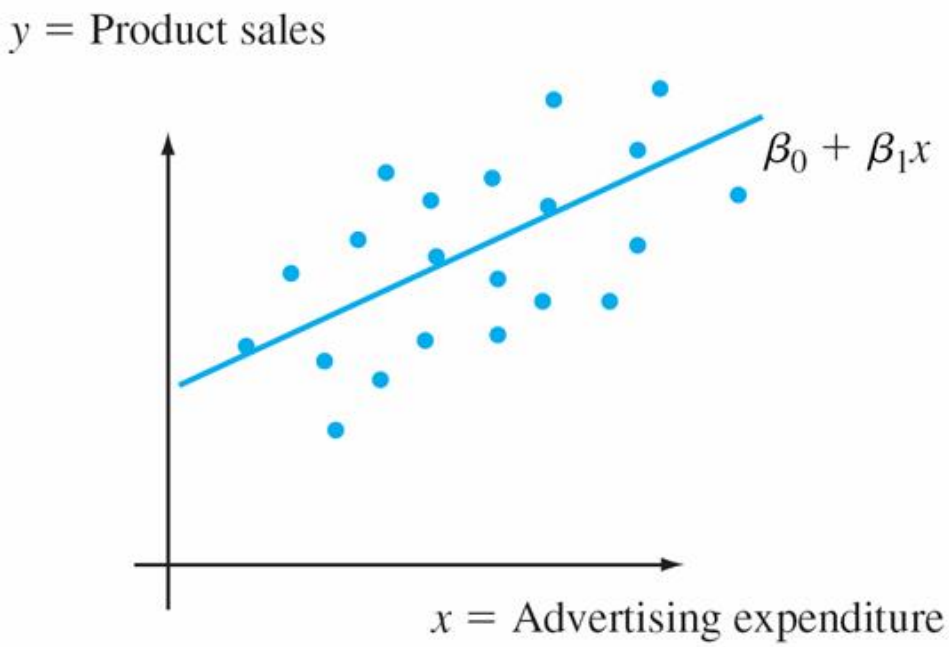
The residual can be thought of as a measure of deviation and we can summarize the notation in the following way:

$$Y_i - \hat{Y}_i = \hat{\epsilon}_i$$

Estimating σ^2 and σ

The parameter σ^2 determines the amount of spread about the true regression line. Two separate examples:





Estimating σ^2 and σ

An estimate of σ^2 will be used in confidence interval (CI) formulas and hypothesis-testing procedures presented in the next two sections.

Many large deviations (residuals) suggest a large value of σ^2, whereas deviations all of which are small in magnitude suggest that σ^2 is small.

Estimating σ^2 and σ

The **error sum of squares** (equivalently, *residual sum of squares*), denoted by SSE, is

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum(y - \hat{y}_i)^2}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$$

(Note that that the homoscedasticity assumption comes into play here.)

Estimating σ^2 and σ

The divisor $n - 2$ in s^2 is the number of degrees of freedom (df) associated with SSE and the estimate s^2.

This is because to obtain s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of 2 df (just as μ had to be estimated in one sample problems, resulting in an estimated variance based on $n - 1$ df in our previous t-tests).

Replacing each y_i in the formula for s^2 by the r.v. Y_i gives the estimator S^2.

It can be shown that the r.v. S^2 is an unbiased estimator for σ^2

Example (variance estimator)

The residuals for the *filtration rate–moisture content data* were calculated previously.

The corresponding *error sum of squares* is

$$SSE = (-.200)^2 + (-.188)^2 + \cdots + (1.099)^2 = 7.968$$

The estimate of σ^2 is then $\hat{\sigma}^2 = s^2 = 7.968/(20 - 2) = .4427$, and the *estimated standard deviation* is

$$\hat{\sigma} = s = \sqrt{.4427} = .665$$

Roughly speaking, .665 is the *magnitude of a typical deviation from the estimated regression line*—some points are closer to the line than this and others are further away.

Estimating σ^2 and σ

Computation of SSE from the defining formula involves much tedious arithmetic, because both the predicted values and residuals must first be calculated.

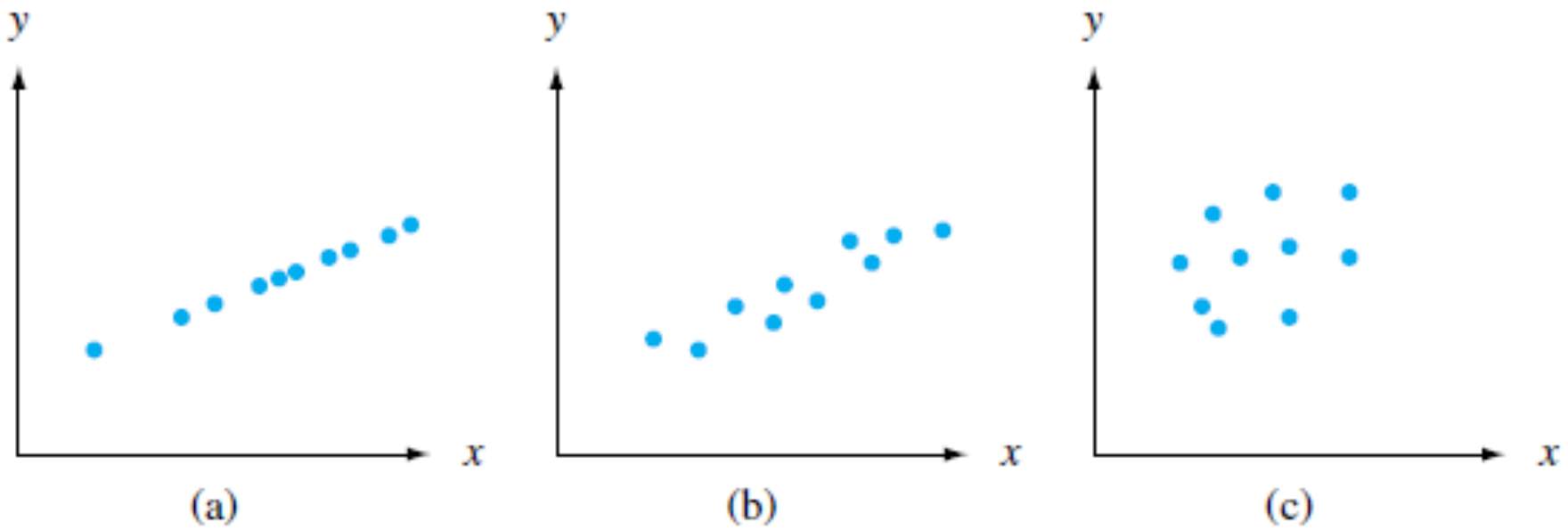
Use of the following **shortcut formula** does not require these quantities.

$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

This expression results from substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into $\sum (y_i - \hat{y}_i)^2$, squaring the summand, carrying through the sum to the resulting three terms, and simplifying.

The Coefficient of Determination

Different variability in observed *y* values:



Using the linear model to explain y variation:
(a) data for which all variation is explained;
(b) data for which most variation is explained;
(c) data for which little variation is explained

The Coefficient of Determination

(a) The points in the first plot all fall exactly on a straight line. In this case, *all (100%) of the sample variation in y can be attributed to the fact that x and y are linearly related* in combination with variation in *x*.

(b) The points in the second plot do not fall exactly on a line, but compared to overall *y* variability, the deviations from the least squares line are small.

It is reasonable to conclude in this case that *much of the observed y variation can be attributed to the approximate linear relationship between the variables* postulated by the simple linear regression model.

(c) When the scatter plot looks like that in the third plot, there is substantial variation about the least squares line relative to overall *y* variation, so *the simple linear regression model fails to explain variation in y by relating y to x*.

The Coefficient of Determination

The *error sum of squares* SSE can be interpreted as a measure of how much variation in *y* is left *unexplained by the model*—that is, how much cannot be attributed to a linear relationship.

In the first plot SSE = 0, and there is no unexplained variation, whereas unexplained variation is small for second, and large for the third plot.

A quantitative measure of the total amount of variation in observed *y* values is given by the **total sum of squares**

$$\text{SST} = S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

The Coefficient of Determination

Total sum of squares is the sum of squared deviations about the sample mean of the observed *y* values – when no predictors are taken into account.

Thus the same number $\overline{y}$ is subtracted from each y_i in SST, whereas SSE involves subtracting each different predicted value $\hat{y}_i$ from the corresponding observed y_i.

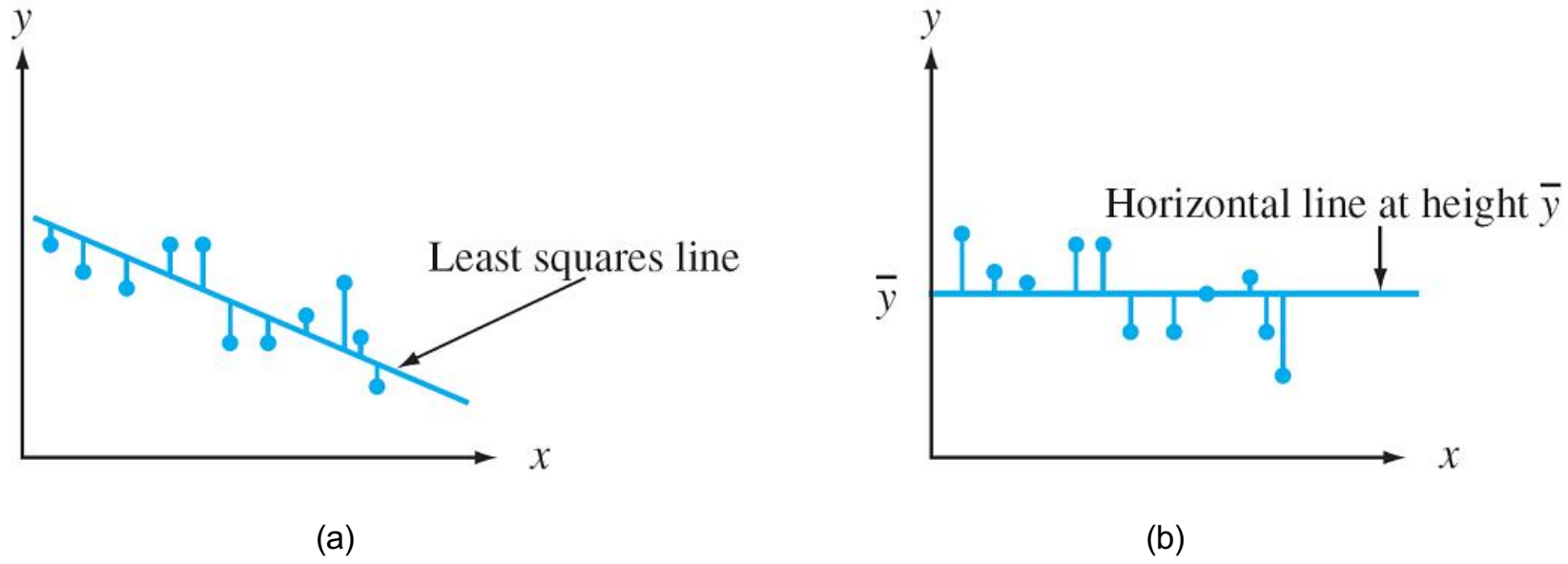
The SST in some sense is *as bad as SSE can get* if there is no regression model (i.e., slope is 0) then

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \Rightarrow \quad \hat{y} = \hat{\beta}_0 + \underbrace{\hat{\beta}_1}_{=0} \overline{x} = \hat{\beta}_0 = \overline{y}$$

Which motivates the definition of the SST.

The Coefficient of Determination

Just as SSE is the sum of squared deviations about the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$, SST is the sum of squared deviations about the horizontal line at height $\overline{y}$ as pictured below:



(a) (b)

Sums of squares illustrated: (a) SSE = sum of squared deviations about the least squares line; (b) SSE = sum of squared deviations about the horizontal line

The Coefficient of Determination

The sum of squared deviations about the least squares line is smaller than the sum of squared deviations about *any* other line, i.e. SSE < SST unless the horizontal line itself is the least squares line.

The ratio SSE/SST is the *proportion of total variation that cannot be explained* by the simple linear regression model, and r^2 = 1 – SSE/SST (a number between 0 and 1) is the proportion of observed *y* variation explained by the model.

Note that if SSE = 0 as in case (a), then $r^2 = 1$.

The Coefficient of Determination

Definition

The **coefficient of determination**, denoted by r^2, is given by

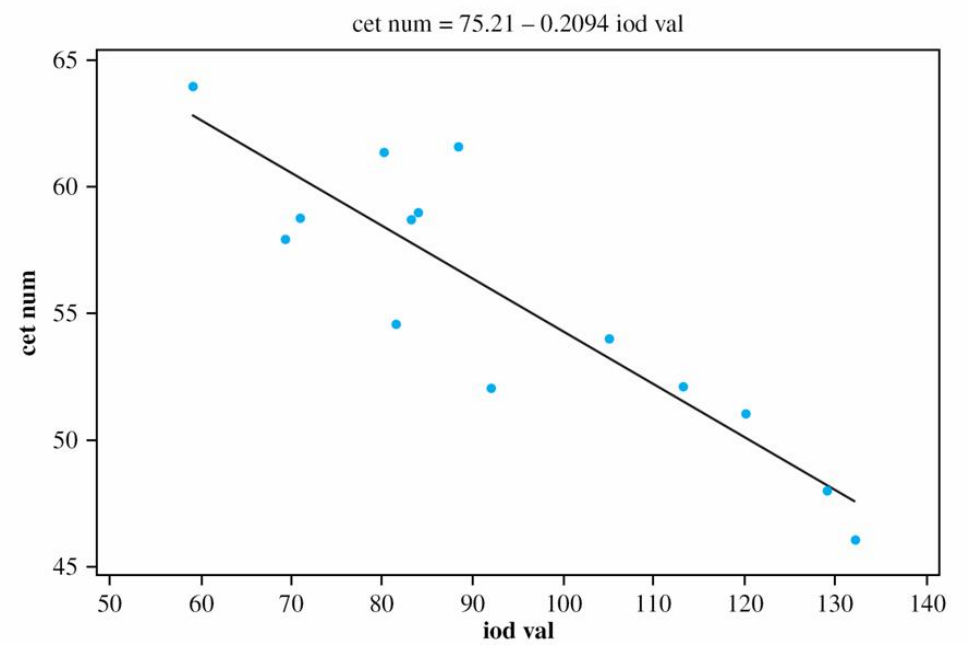
$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{S_{yy}}$$

It is interpreted as the *proportion of observed y variation that can be explained by the simple linear regression model* (attributed to an approximate linear relationship between *y* and *x*).

The higher the value of r^2, the more successful is the simple linear regression model in explaining *y* variation.

Example

The scatter plot of the iodine value–cetane number data in the previous example implies a reasonably high r^2 value.



Scatter plot for Example 4 with least squares line superimposed, from Minitab

Example

cont'd

The coefficient of determination for the previous example is then

$$r^2 = 1 - SSE/SST = 1 - (78.920)/(377.174) = .791$$

That is, 79.1% of the observed variation in cetane number is attributable to (can be explained by) the simple linear regression relationship between cetane number and iodine value.

The Coefficient of Determination

The coefficient of determination can be written in a slightly different way by introducing a third sum of squares—**regression sum of squares,** SSR—given by

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = SST - SSE.$$

Regression sum of squares is interpreted as the amount of total variation that *is* explained by the model.

Then we have

$$r^2 = 1 - SSE/SST = (SST - SSE)/SST = SSR/SST$$

the ratio of explained variation to total variation.

Inferences About the Slope Parameter β_1

In virtually all of our inferential work thus far, the notion of sampling variability has been pervasive.

Properties of sampling distributions of various statistics have been the basis for developing confidence interval formulas and hypothesis-testing methods.

Same idea as before: The value of any quantity calculated from *sample data* (which is random) will vary from one sample to another.

Inferences About the Slope Parameter β_1

The **estimators** are:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} = \sum c_i Y_i$$

That is, $\hat{\beta}_1$ is a linear function of the independent rv's $Y_1, Y_2, \ldots, Y_n$, each of which is normally distributed.

Similarly, we have the estimators:

$$\hat{\beta}_0 = \frac{\sum Y_i - \hat{\beta}_1 \sum x_i}{n}$$

And,

$$\hat{\sigma}^2 = S^2 = \frac{\sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum x_i Y_i}{n - 2}$$

Inferences About the Slope Parameter β_1

Invoking properties of a linear function of random variables as discussed earlier, leads to the following results.

1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 (the distribution of $\hat{\beta}_1$ is always centered at the value of β_1, which is unknown).

2. The variance and standard deviation of β_1 are

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

where $S_{xx} = \Sigma(x_i - \bar{x})^2$ and σ is the (unknown) true st. dev.

Inferences About the Slope Parameter β_1

Replacing σ by its estimate s gives an estimate for $\sigma_{\hat{\beta}_1}$ (the estimated standard deviation, i.e., estimated standard error, of $\hat{\beta}_1$):

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$$

This estimate can also be denoted by $\hat{\sigma}_{\hat{\beta}_1}$.

(Recall $s^2 = SSE/n\text{-}2$)

3. The estimator $\hat{\beta}_1$ has a <u>normal</u> distribution (because it is a linear function of independent normal r.v.'s).

Inferences About the Slope Parameter β_1

NOTE:

- x_i values that are quite spread out = estimator with a low standard error.
- x_i all close to one another = highly variable estimator.

If the x_i's are spread out too far, a linear model may not be appropriate throughout the range of observation.

Inferences About the Slope Parameter β_1

Theorem

The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

has a *t* distribution with $n - 2$ df (since $\sigma \approx s$).

A Confidence Interval for β_1

As in the derivation of previous CIs, we begin with a probability statement:

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate β_1 and substitution of estimates in place of the estimators gives the CI formula.

A 100(1 – α)% **CI for the slope β_1** of the true regression line is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

Example

Variations in clay brick masonry weight have implications not only for structural and acoustical design but also for design of heating, ventilating, and air conditioning systems.

The article "Clay Brick Masonry Weight Variation" (*J. of Architectural Engr.*, 1996: 135–137) gave a scatter plot of y = mortar dry density (lb/ft^3) versus x = mortar air content (%) for a sample of mortar specimens, from which the following representative data was read:

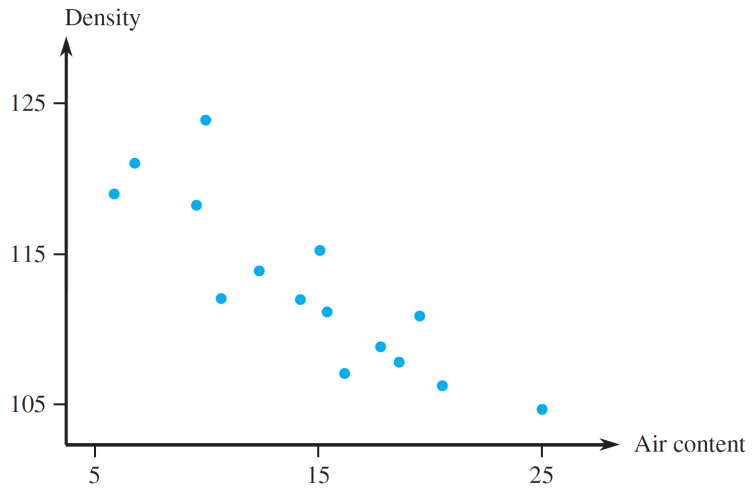
x	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0	15.3
y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1	111.3

x	16.2	17.8	18.7	19.7	20.6	25.0
y	107.2	108.9	107.8	111.0	106.2	105.0

Example

cont'd

The scatter plot of this data in Figure 12.14 certainly suggests the appropriateness of the simple linear regression model; there appears to be a substantial negative linear relationship between air content and density, one in which density tends to decrease as air content increases.



Scatter plot of the data from Example 11

Figure 12.14

Example

cont' d

The values of the summary statistics required for calculation of the least squares estimates are

$\Sigma x_i = 218.1 \quad \Sigma y_i = 1693.6 \quad \Sigma x_i y_i = 24{,}252.54 \quad \Sigma x_i^2 = 3577.01$

$\Sigma y_i^2 = 191{,}672.90; \; n = 15$

What is r^2 and how is it interpreted?

What is the 95% confidence interval for the slope?

Hypothesis-Testing Procedures

The most commonly encountered pair of hypotheses about β_1 is H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. When this null hypothesis is true, $\mu_{Y \cdot x} = \beta_0$ (independent of x). Then knowledge of x gives no information about the value of the dependent variable.

Null hypothesis: H_0: $\beta_1 = \beta_{10}$

Test statistic value: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$ ("t ratio")

Hypothesis-Testing Procedures

Alternative Hypothesis	**Alternative Hypothesis**
H_a: $\beta_1 > \beta_{10}$	$t \geq t_{\alpha,n-2}$
H_a: $\beta_1 < \beta_{10}$	$t \leq -t_{\alpha,n-2}$
H_a: $\beta_1 \neq \beta_{10}$	either $t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$

A *P*-value based on $n - 2$ can be calculated just as was done previously for *t* tests.

If H_0: $\beta_1 = 0$, then the test statistic is the ***t* ratio** $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

Regression in R.

Inference Concerning Mean of Future Y

Let x^* denote a specified value of the independent variable x.

Once the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have been calculated, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ can be regarded either as a point estimate of $\mu_{Y \cdot x^*}$ (the *expected* or *true average value of Y when* $x = x^*$) or as *a prediction of the Y value that will result from a single observation made when* $x = x^*$.

Inference Concerning Mean of Future Y

The estimate of $\mu_{Y \cdot x^*}$ is random, so we can develop a CI for $\mu_{Y \cdot x^*}$ and a *prediction interval* (PI) for a single *Y* value. **What is the difference?**

Before we obtain sample data, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are subject to sampling variability—they are both statistics whose values will vary from sample to sample.

Suppose, for example, that the *true* $\beta_0 = 439$ and $\beta_1 = 0.05$. Then a first sample of (*x*, *y*) pairs might give $\hat{\beta}_0 = 439.35$, $\hat{\beta}_1 = 0.048$; a second sample might result in $\hat{\beta}_0 = 438.52$, $\hat{\beta}_1 = 0.051$; and so on.

Inference Concerning Mean of Future Y

It follows that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ itself *varies* in value from sample to sample – **it is a random variable**.

If the intercept and slope of the population line are the values 439 and 0.05, respectively, and suppose x^* = 5kgs, then this statistic is trying to *estimate* the *true* value which is:

$$439 + 0.05(5) = 439.25 = \mu_{Y \cdot x^*}$$

Then the estimate from a first sample might be

$$439.35 + 0.048(5) = 439.59,$$

from a second sample it might be

$$438.52 + 0.051(5) = 438.775 \text{ , and so on.}$$

Inference Concerning Mean of Future Y

Inferences about the mean *Y*-value
$\hat{\beta}_0 + \hat{\beta}_1 x^*$ will be based on properties of the sampling distribution of the statistic $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

Substitution of the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{\beta}_0 + \hat{\beta}_1 x^*$ followed by some algebraic manipulation leads to the representation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as a linear function of the Y_i's:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] Y_i = \sum_{i=1}^{n} d_i Y_i$$

The coefficients $d_1, d_2, \ldots, d_n$ in this linear function involve the x_i's and x^*, all of which are fixed.

Inference Concerning Mean of Future Y

Application of the rules to this linear function gives the following properties.

Proposition

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ where x^* is some fixed value of x. Then

1. The expectation of $\hat{Y}$ is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

Thus $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ is an *unbiased estimator* for $\beta_0 + \beta_1 x^*$ (i.e., for $\mu_{Y \cdot x^*}$).

Inference Concerning Mean of Future Y

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma^2_{\hat{Y}} = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum x_i^2 - (\sum x_i)^2/n}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

And the standard deviation $\sigma_{\hat{Y}}$ is the square root of this expression. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$, denoted by $s_{\hat{Y}}$ or $s_{\hat{\beta}_0+\hat{\beta}_1 x^*}$, results from replacing σ by its estimate s (recall $s^2 = SSE/n\text{-}2$):

$$s_{\hat{Y}} = s_{\hat{\beta}_0+\hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y}$ has a **normal distribution**.

Inference Concerning Mean of Future Y

The variance of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$ in either direction.

Thus the estimator of $\mu_{Y \cdot x^*}$ is more precise when x^* is near the center of the x_i's than when it is far from the values at which observations have been made. This will imply that both the CI and PI are narrower for an x^* near $\bar{x}$ than for an x^* far from $\bar{x}$.

Distribution of Future Y

Just as inferential procedures for β_1 were based on the *t* variable obtained by standardizing β_1, a *t* variable obtained by standardizing $\hat{\beta}_0 + \hat{\beta}_1 x^*$ leads to a CI and test procedures here.

Theorem

The variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - E(\hat{Y})}{S_{\hat{Y}}} = \frac{\hat{Y} - Y}{S_{\hat{Y}}}$$

has a *t* distribution with $n - 2$ df.

Confidence Interval for Future Y

A probability statement involving this standardized variable can now be manipulated to yield a confidence interval for

$\mu_{Y \cdot x^*}$

A **100(1 – α)% CI for the expected value of *Y* when $x = x^*$**, is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2,\, n-2} \cdot s_{\hat{Y}}$$

This CI is centered at the point estimate for $\mu_{Y \cdot x^*}$ and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator on which the point estimate is based.

Example: CI for $Y|X=x$ based on regression

Corrosion of steel reinforcing bars is the most important durability problem for reinforced concrete structures.

Carbonation of concrete results from a chemical reaction that also lowers the pH value by enough to initiate corrosion of the rebar.

Representative data on

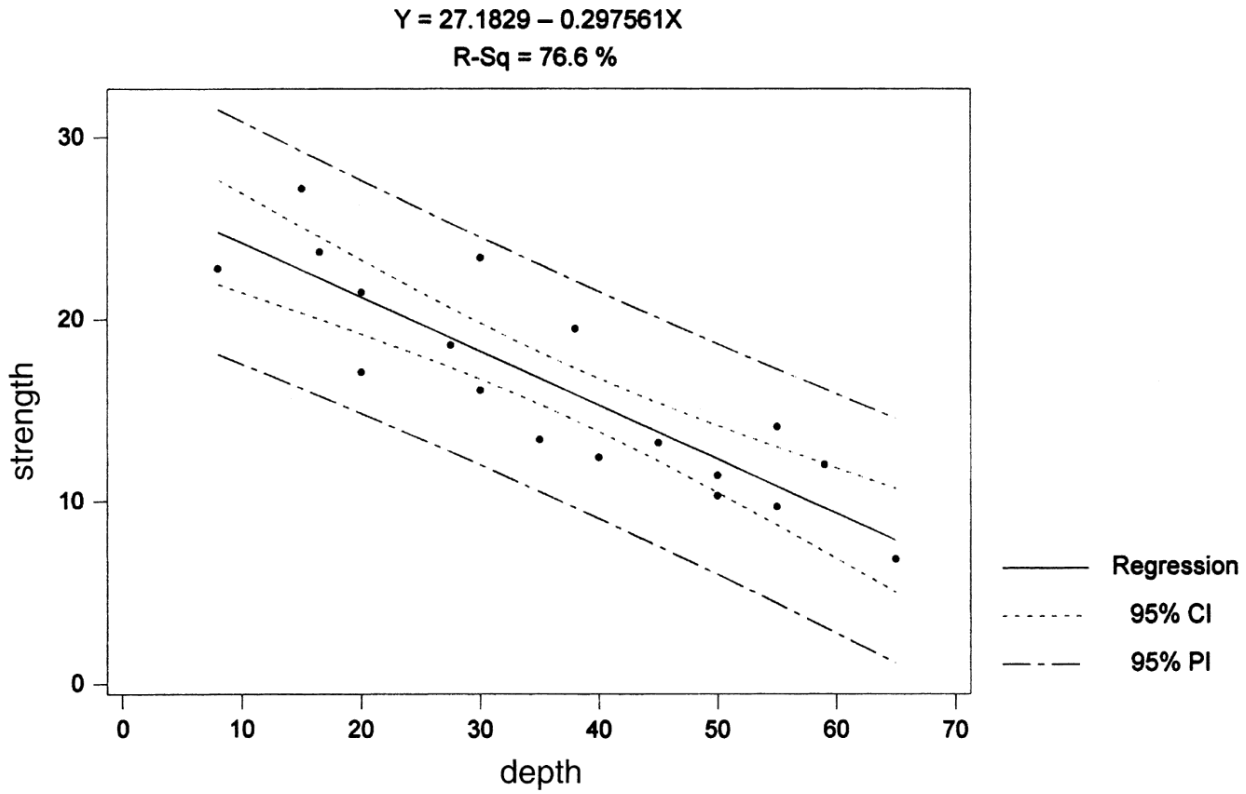
x = carbonation depth (mm)

and y = strength (MPa) for a sample of core specimens taken from a particular building

follows on the next slide

Example: CI for $Y|X=x$ based on regression cont' d

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4
x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8



Example: CI for $Y|X=x$ based on regression cont' d

Let's now *calculate a 95% confidence interval for the mean strength for all core specimens having a carbonation depth of 45.*

A Prediction Interval for a Future Value of *Y*

Rather than calculate an interval estimate for $\mu_{Y \cdot x^*}$, an investigator may wish to obtain a range or *an interval of possible values of Y* associated with some future observation when the independent variable has value x^*.

Consider, for example, relating vocabulary size *y* to age of a child *x*. The CI with $x^* = 6$ would provide a range that covers with 95% confidence *the true average vocabulary size for all 6-year-old children*.

Alternatively, we might wish *an interval of plausible values for the vocabulary size of a particular 6-year-old child*. How can you tell that a child is "off the chart" for example?

A Prediction Interval for a Future Value of *Y*

A confidence interval refers to a *parameter*, or population characteristic, whose value is *fixed but unknown* to us.

In contrast, a future value of *Y* is not a parameter but instead a random variable; for this reason we refer to an *interval of plausible values for a future Y* as a **prediction interval** rather than a confidence interval.

Determining a prediction interval for *Y* requires that we model the error involved in the prediction of the *Y* variable.

A Prediction Interval for a Future Value of *Y*

The error of prediction is $Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, *i.e.* a difference between two random variables. Because the future value *Y* is independent of the observed Y_i's, we have

$$\begin{aligned}
\text{variance of prediction error} &= V[Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)] \\
&= V(Y) - V(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\
&= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right] \\
&= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right]
\end{aligned}$$

A Prediction Interval for a Future Value of *Y*

Furthermore, because $E(Y) = \beta_0 + \beta_1 x^*$ and expectation of $\hat{\beta}_0 + \hat{\beta}_1 x^* = \beta_0 + \beta_1 x^*$, the expected value of the prediction error is $E(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)) = 0$.

It can then be shown that the standardized variable

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} = \frac{(Y - \hat{Y}) - 0}{S_{Y-\hat{Y}}} = \frac{(Y - \hat{Y}) - E(Y - \hat{Y})}{S_{Y-\hat{Y}}}$$

has a *t* distribution with $n - 2$ df.

A Prediction Interval for a Future Value of *Y*

Manipulating to isolate *Y* between the two inequalities yields the following interval.

A 100(1 – α)% **PI for a future *Y* observation to be made when $x = x^*$** is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}}$$

$$= \hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{Y}}}$$

A Prediction Interval for a Future Value of *Y*

The interpretation of the prediction level 100(1 – α)% is similar to that of previous confidence levels—if is used repeatedly, in the long run the resulting interval will actually contain **the observed *y* values** 100(1 – α)% of the time.

Notice that the 1 underneath the initial square root symbol makes the PI wider than the CI, though the intervals are both centered at $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Also, as $n \to \infty$, the width of the CI approaches 0, whereas the width of the PI does not (because even with perfect knowledge of β_0 and β_1, there will still be randomness in prediction).

Example: PI for $Y|X=x$ based on regression

Return to the carbonation depth-strength data example and *calculate a 95% PI for a strength value that would result from selecting a single core specimen whose depth is 45 mm.*

Residuals and Standardized Residuals

The **standardized residuals** are given by

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \quad i = 1, \ldots, n$$

If, for example, a particular standardized residual is 1.5, then the residual itself is 1.5 (estimated) standard deviations larger than what would be expected from fitting the correct model.

Diagnostic Plots

The basic plots that many statisticians recommend for an assessment of *model validity* and usefulness are the following:

1. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis

2. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis

3. $\hat{y}_i$ on the vertical axis versus y_i on the horizontal axis

4. A histogram of the standardized residuals

Diagnostic Plots

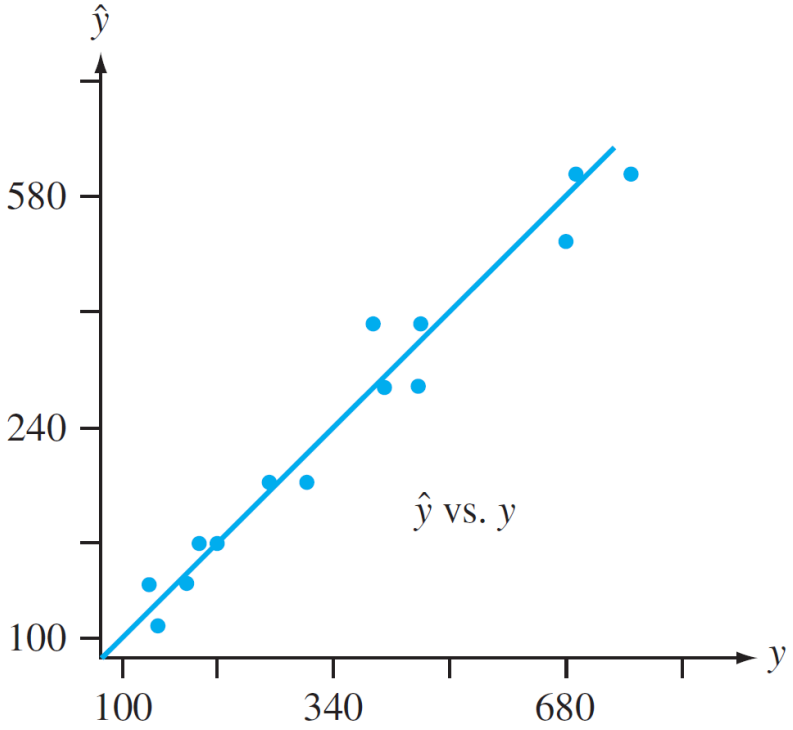
Plots 1 and 2 are called **residual plots** (against the independent variable and fitted values, respectively), whereas Plot 3 is fitted against observed values.

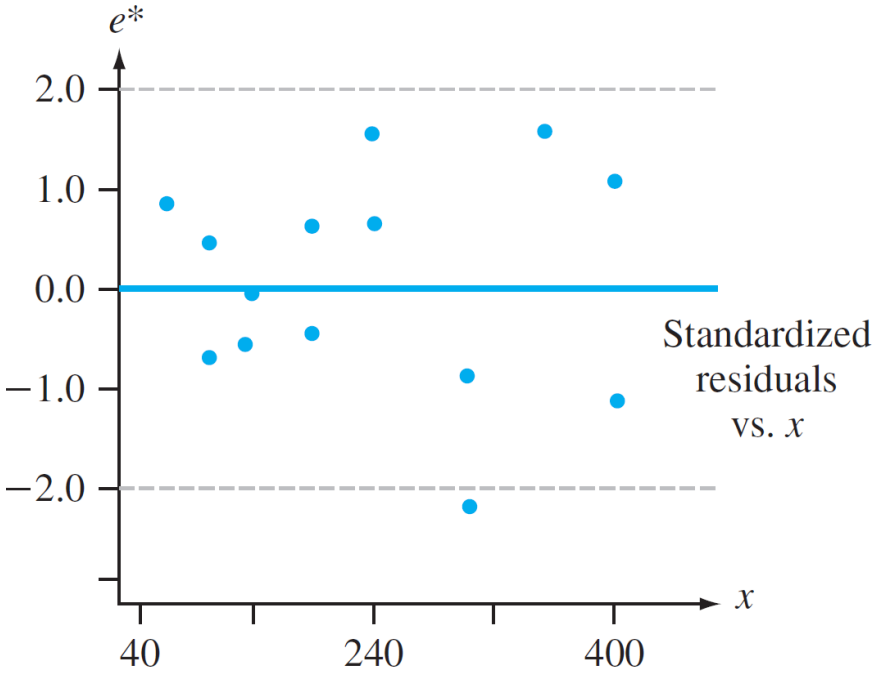
Provided that the model is correct, *neither residual plots should exhibit distinct patterns*.

The residuals should be randomly distributed about 0 according to a normal distribution, so all but a very few standardized residuals should lie between –2 and +2 (i.e., all but a few residuals within 2 standard deviations of their expected value 0).

If Plot 3 yields points close to the 45-deg line [slope +1 through (0, 0)], then the estimated regression function gives accurate predictions of the values actually observed.

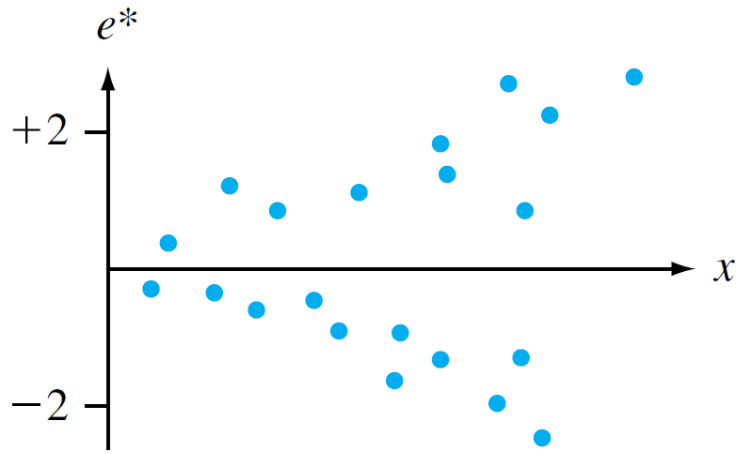
Example (Plot Type #2 and #3)





Heteroscedasticity

The residual plot below suggests that, although a straight-line relationship may be reasonable, the assumption that $V(Y_i) = \sigma^2$ for each i is of doubtful validity.



Using advanced methods like weighted LS (WLS), or more advanced models, is recommended for inference.